DOCUMENT RESUME

ED 195 302 JC 800 626

AUTHOR O'Reilly, Patrick A.: And Others
TITLE Evaluating Performances. Implementing Competency Based Education in Community Colleges, 4.
INSTITUTION Virginia Polytechnic Inst. and State Univ., Blacksburg.
SPONS AGENCY Michigan State Dept. of Education, Lansing. Higher Education Management Services.
PUB DATE [80]
NOTE 31p.: For related documents, see JC 800 623-632.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Community Colleges: *Competency Based Education: Evaluation Methods: Grades (Scholastic): *Performance Tests: *Student Evaluation: Test Construction: Test Format: *Testing: Testing Problems: Test Validity: Two Year Colleges: Vocational Education

ABSTRACT

After introductory material examining student skepticism toward testing and outlining the major problems encountered by instructors in test construction and grading, this monograph explores the optimum use of testing and performance evaluations in competency-based vocational education programs. Background information is then presented, including as assessment of the differences between "testing" and "evaluation," an outline of the major purposes of evaluation, and a description of commonly used evaluation techniques and testing instruments. Problems in evaluation are then discussed, in relation to the development of valid test measures, the statistical reliability of testing instruments, and the difficulty of assessing affective skills. The monograph then provides suggestions for (1) assuring the relevance of evaluations by using appropriate testing techniques, establishing a realistic testing environment, and basing evaluations on articulated course objectives and actual job requirements: (2) insuring that evaluation results are useful to both students and instructors: (3) simplifying evaluations by using efficient procedures for instrument construction and correction: (4) avoiding common measurement problems such as identifiable answer patterns: (5) converting the results of competency-based evaluations to traditional grades: and (6) using alternatives to traditional grades. (JP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED195302
C
B
E
JC 800 626

EVALUATING PERFORMANCES

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

James H. Folkening

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PATRICK A. O'REILLY
DANIEL E. VOGLER
F. MARION ASCHE

This series of monographs includes the following:

1. Using Performance Goals
2. Using Performance Objectives
3. Developing a Performance Syllabus
4. Evaluating Performances
5. Motivating Performance Learning
6. Developing and Using Performance Modules
7. Addressing Affective Performance
8. Remediating and Enriching Performance Learning
9. Managing a Performance On The Job
10. Maximizing Teaching/Learning

This is one of ten monographs in a series relating to IMPLEMENTING COMPETENCY BASED EDUCATION IN COMMUNITY COLLEGES developed at Virginia Polytechnic Institute and State University. This project was developed under a grant to St. Clair County Community College and was funded by the Michigan Department of Education, Higher Education Management Services, Occupational Education Unit. Project #33B9-7447-2.

Evaluating Performance

Testing and evaluation is an integral part of the educational process. Although there has been considerable discussion regarding the value and meaning of testing and evaluation, no drastic reduction in the use of testing appears imminent. Students and instructors generally seem to have accepted this fact, perhaps with some reluctance. Both groups have questions and concerns about the evaluation process. While these questions are asked from different points of view, they are still useful and should serve as the basis for a practical discussion of evaluation.

Students' Concerns

Students' questions are most often concerned with the evaluation process after the instrument has been constructed. This is understandable since this is the part of the process where they have involvement. Examples of the types of questions students frequently ask about testing and evaluation are:

1. How important is the test?
2. What will the test cover?
3. How are grades determined?
4. What will the test be like?
5. How many times can the test be taken?
6. What is the minimum acceptable performance?
7. What happens if I fail?
8. Will skills and knowledge previously evaluated be reevaluated later?
9. How does evaluation relate to what is needed on the job?
10. Why don't students have input concerning evaluation?

The above list is neither comprehensive nor sequenced by any importance. However, all instructors would probably agree that the first eight questions are more frequently asked by students than are questions of the type represented by the last two.

Students' questions about testing and evaluation provide some insight into how they feel about the process. One of the most evident feelings is that they are threatened by the process. Often rather than viewing evaluation as a challenge, a chance to demonstrate their knowledge or skill, they see it as being punitive. In general, students

exhibit a feeling of powerlessness with regard to the evaluation process. Students see themselves as being at the mercy of the process. Their questions reflect a feeling that the process is vague and unspecified. It is a game of academic roulette. They also question the value or relevance of the process. It is often seen as a way of putting students into categories that have little to do with what they know or can do.

Overall the students' views of tests and evaluation are exhibited as an attitude. This attitude might be best described by paraphrasing Vince Lombardi, "Passing isn't the most important thing - it's the only thing." More emphasis is placed on obtaining a passing grade than on learning. This attitude affects how students view both the content and instruction and makes the instructor's job more difficult.

Instructors' Concerns

Instructors also raise questions about tests and evaluation. In some instances these questions closely parallel or resemble those asked by students. In general, instructors' questions focus on how evaluation should be planned, conducted, and used. Many appear to be slightly uneasy about their role as an evaluator and the designer of evaluation. Examples of the type of questions frequently asked by instructors are:

1. Can standardized tests be used in a competency-based program?
2. Is only skill performance evaluated in a competency-based program?
3. How important are teacher-made tests?
4. What should be evaluated?
5. How can the students' ability to apply knowledge be measured?
6. How can evaluation be made realistic?
7. How can competency-based achievement be translated into letter grades?
8. How well can students' knowledge be measured?
9. How can the evaluation of students help improve instruction?
10. How can the evaluation of students be made easier?

Again, the above list is neither comprehensive nor sequenced. It is an example of the types of questions instructors often ask.

The questions instructors ask seem to be related to several major issues. First is the issue of what to evaluate and how. Second is the

issue of improving evaluation. How the results of evaluation can and should be used is a third issue. And finally, the fourth and overriding issue is how effective evaluation can be accomplished in a reasonable amount of time. The fourth issue, effective evaluation given time constraints, probably is among the most difficult problems faced by the average instructor. All other issues must be addressed with this issue in mind. Unfortunately, instructors often view evaluation as a time-consuming procedure that primarily affects students. It is even more unfortunate that this attitude is supported by rationalization that standardized tests have no more effect on students' futures than do teacher-made tests. Few beliefs could be less true!

Perspective

It is interesting and important to note that while students and instructors appear to have distinct concerns about testing and evaluation, they are really concerned about many of the same things. Students are concerned about what is required of them; instructors are concerned about what should be required. Students wonder how evaluation relates to job requirements; so do instructors. Students want their skill and knowledge measured accurately; instructors wonder how to do so. Students see evaluation as taking up instructional time; so do instructors. Rather than being adversaries on the issues surrounding tests and evaluation, students and their instructors should be allies.

The remainder of this monograph is intended to provide at least partial answers to the questions previously identified. Although the remaining materials may at times verge on the theoretical or academic, such information has been kept to a minimum.

Testing and Evaluation

The purpose of this section is to provide a brief review of testing and evaluation. It is intended to identify major principles and problems in the evaluation process as well as to establish a solid foundation from which the suggestions in the following section can be built.

Although they are often considered to be the same, testing and evaluation are distinctly separate activities. Testing is part of a larger activity called **measurement.** Measurement is the process of determining how much exists, how big something is, or how far something is from something else. It is a procedure in which a known is compared to an unknown. Measurement may be based on either a formal or informal standard. For example, the lengths of two lines

could be described as six inches and three inches or as one's being twice as long as the other. The extent to which formal and standardized measurement is necessary depends largely upon why the measurement is being made.

By definition then, measurement is an objective process. The accuracy of various instruments used may vary and thus the outcome of the measurement process may be more or less accurate. However, the outcome alone does not include any judgment about how good or bad whatever was measured might be. To say that a two by four is thirty-two inches long is merely a description. However, if a forty-inches long two by four is needed, comparing what is needed to what is available is an indication of how useful it might be. Likewise, to say that John Jones scored ten points on a test is, in and of itself, just a description of John's performance. To even know that John scored ten points out of fifty is still an objective description of his performance. Without additional information it is impossible to make a determination about how good or bad his performance was.

On the other hand, evaluation is a subjective process. In order for evaluation to be possible, measurement of some kind must have already been conducted. Evaluation is the process of making some judgments regarding how good or bad something is. It is only possible to make evaluative judgements if there is some standard to compare against. The extent to which the standard is clearly specified and measurable is the major factor determining how subjective the evaluation will be. For example, indicating that a passing grade will be assigned to all students obtaining a good score on the test leaves much more room for interpretation than stating that a passing grade will be assigned for all test scores of 60 percent or higher. Another factor affecting the evaluation process is the appropriateness of the standard being used. This issue will be discussed in some detail later.

Purposes of Evaluation

Many instructors view the need to establish a bias for assigning grades as the main incentive for evaluating student performance. Almost all students believe this to be the only purpose of evaluation. While this is a valid purpose, it is only one of many. Instructors who use evaluation for other purposes are generally more effective and often construct better instruments. Moreover, using a test for multiple purposes makes good sense in light of the amount of time an instructor invests in planning and conducting a good evaluation. It would be impossible for a single evaluation to serve all of the pur-

poses discussed here; however, with little extra effort two or more could be served.

Assessing achievement. The most common purpose of evaluation is to assess student achievement; that is, to measure how much they know and how well they can do something. It is essential to determining how much students have learned. This is the purpose of evaluation most closely associated with the assignment of grades.

Assessing knowledge or skill. Often it is desirable to ascertain the amount students know or their level of skill for diagnostic or prognostic purposes. Evaluation can be used to identify what students already can do or know as well as what they can't do or don't know. This information can be used to restructure instruction (diagnostic) or make judgments about the type job at which the student will be most successful (prognostic).

Stimulating or reinforcing learning. Evaluation can be used to stimulate student learning. This has always been one of the underlying assumptions for measuring student achievement. The most effective way to stimulate learning is to use pre-instructional measures to give students an idea of what will be taught and what is important. Evaluation provides feedback to students about how well they did. However, a conscious effort must be made by the instructor if the feedback is to provide positive and constructive reinforcement.

Assessing program success. While one generally associates program objectives with special programs, traditional programs also have objectives. Although the objectives of traditional programs often are not explicitly stated, they should be. The objectives of a program are distinct from the objectives set for students in the program. An example of a program objective is: "At the conclusion of the program, 90 percent of the students will be able to achieve a score of 75 percent on the Short Occupational Knowledge Test (SOKT) for auto mechanics." The success of a program in achieving such an objective could be easily assessed through evaluation. Accountability will increasingly demand more evaluative information about programs.

Assessing teaching. Most evaluations of students can provide some information relative to teaching effectiveness. In some situations this may be the primary purpose of evaluation. For example, an instructor may wish to compare the achievement of students using self-instructional materials with those receiving traditional instruction.

The situation need not be as clear-cut, however. Even when different instructional methods are not being compared, the results of evaluation provide feedback that is useful in improving teaching.

Summary. There may be additional purposes for conducting an evaluation. As has been discussed, several purposes may be served by one evaluation. Care should be taken to insure the evaluation is conducted in a manner compatible with the purposes and that the purposes themselves are compatible.

Major Steps in Evaluation

The evaluation process can be broken into several major steps. The following is a suggested list of these steps.

1. Determine the purpose of the evaluation
2. Determine what to evaluate
3. Determine measurement procedures
4. Determine evaluation criteria
5. Construct or select the measurement/evaluation
6. Conduct the evaluation
7. Determine the results
8. Evaluate the evaluation
9. Determine areas needing improvement
10. Implement appropriate changes

It is not always necessary to formally go through every step. However, to at least do so on an informal basis emphasizes two important points. First, evaluation is a process. Second, effective evaluation is most often achieved by careful planning.

Types of Evaluation

There are a number of ways in which evaluation can be discussed. One way is to look at the various types of evaluations that can be conducted. Many of these types can be combined and any particular evaluation might include several.

Formal or informal. Formal evaluation, generally planned and scheduled, requires the student to perform in a specific way. Tests given in the classroom are examples of formal evaluation. It most often represents an attempt to measure the students' knowledges and skills when they know evaluation is taking place and are performing at their best. Informal evaluation is often an on-going process and frequently based on observations. It is an attempt to determine how students perform on a day-to-day basis; that is, how they typically perform. The most common examples are checklists, rating scales,

and anecdotal notes. It should be noted that formal evaluation is more likely to be based on objective measures while informal evaluation is often based on feelings, personalities, and chance observation. Likewise, checklists are more objective than either rating scales or anecdotal notes. Both formal and informal evaluation serve useful purposes. However, the utility of informal evaluation is dependent upon the instructor's awareness of the problems involved and his maintaining objectivity.

Objective or subjective. Objectivity and subjectivity have more often been associated with the types of test items used than with evaluation. In actuality, the situation is not that clear-cut. Evaluation based upon clearly specified criteria known to the learner is generally objective while evaluation based upon vague and hidden criteria is subjective. The objectivity of the evaluation process encompasses the selection of what will be measured as well as how. One of the purposes of competency-based education is to reduce the subjectivity of evaluation by specifying what is important (standards) for success. Subjective evaluation places the burden on the instructor for justifying the evaluation.

Norm or criterion referenced. Norm-referenced evaluation attempts to answer the question of how well each student performed when compared to the other students being evaluated. Most evaluation in education is norm referenced. On the other hand, criterion-referenced evaluation attempts to ascertain whether the student can perform well enough to meet a specified standard. Competency-based education generally uses the principle of criterion-referenced evaluation. When a norm-referenced approach is used, the students are forced to compete with classmates or other students to achieve success. A criterion-referenced approach more nearly forces the student to compete with oneself. A major drawback of norm-referenced evaluation is that standards change from class to class. What constitutes "A" performance in one class may only be "B" performance in another. With criterion-referenced evaluation, the level of performance required for a satisfactory outcome remains constant.

Formative or summative. Formative evaluation, conducted during the evaluation process, is concerned with program efficiency. When evaluating students, it is designed to provide feedback on how students are progressing toward the final objective. For example, evaluations of student achievement on intermediate or enabling objectives is formative in nature. Summative evaluation is concerned

with the final product. In program evaluation it is a measure of effectiveness; that is, whether the program accomplished what was intended or desired. Evaluating program graduates in terms of the terminal performance objectives of the course is summative evaluation. On the whole, formative evaluation is designed to determine how one is doing while there is still time for making changes. Summative evaluation is designed to describe how well one did upon completion.

Entry or exit. Entry evaluation is an attempt to measure how much the student knows, can do, or is capable of, relative to the content of the program before instruction begins. This information can also be used to determine who should be allowed to enter the program. Exit evaluation is an example of summative evaluation. As already discussed, its purpose is to measure what students know or can do at the conclusion of instruction. The difference between what they knew upon entry and what they know upon exit is how much was learned.

Ideally, students should be assigned grades based upon what was learned, not what they know prior to instruction. It is often more difficult to implement such a procedure than to support it.

Individual or group. At times it is advantageous to evaluate students on an individual basis, especially when one student is having unusual difficulty. More often, however, instruction is self-paced or the student's ability to perform a skill sequence is the object of evaluation. If the procedure or process, rather than the product, is to be evaluated, then individual evaluation is required. Group evaluation is more common. The fact that a number of students are being evaluated concurrently makes the procedure more time efficient. While individual evaluation is more time consuming, it often better meets the student's needs and should not be eliminated as a possibility.

Types of Tests

Measurement instruments (tests) can also be discussed according to type. Again, the types of classifications are such that a single instrument could have characteristics from each of the comparisons below.

Instructor-made or standardized. Students take far more instructor-made than standardized tests. Instructor-made tests can be as good as or better than standardized instruments. Their major advantage is that they are specifically constructed based upon first-

hand knowledge of the students involved and what is taught. The major disadvantage is that, although they can be as well constructed as standardized instruments, they often are not. Standardized tests are best suited for comparing the performance of a class or group to a similar national or regional group.

The major advantage of standardized instruments is that they have usually been carefully constructed and checked. Additionally, instructors can reduce the time required for preparation and scoring. They are not without disadvantages, however. Primary among these are (a) they are constructed to evaluate the average student based upon what is perceived to be the usual or important subject matter, and (b) the cost associated with obtaining such instruments is high. Both instructor-made and standardized tests are useful and can be used in a competency-based program.

Pencil and paper as performance. Pencil and paper tests are almost always intended to measure knowledge. There are exceptions, such as in drafting. Moreover, pencil and paper tests can be constructed to approximate or estimate skill performance. Their major and most common use, however, is to measure how much technical-related knowledge the students have learned. They are easily adapted to either group or individual evaluation. Performance tests are skill measures. They can measure either the skill with which students perform a particular task, the quality of the product, or both. They can be time consuming to use, especially if individual evaluation is required. Such is often the case when the process is a better indication of skill than the product.

Timed or power. Quite simply, timed tests have a time limit. The student's achievement is at least partially based upon how much is accomplished within the allowed time. Power tests have no time limit and the amount of time the student takes does not affect the outcome in any way. Therefore, timed tests measure the efficiency of the student's performance while power tests measure how much the student knows or can do. Time is often a factor in performing certain skills and is more often associated with performance tests.

Objective or subjective. Traditionally objective tests have been defined as those employing only true-false, matching, or multiple-choice types of test items. Subjective tests have been described as those in which the judgment or opinion of the scorer can affect the outcome. Most commonly cited are short-answer and essay-type

items. The objectivity of such items can be increased if they are carefully constructed for a specific purpose and if evaluation guidelines or criteria are specified during construction. The use of subjective tests and test items has been widely discouraged. However, they are useful when properly constructed. Although it is difficult and time consuming to construct either a good objective or subjective instrument, scoring a subjective instrument is often more difficult.

Problems in Evaluation

There are a number of problems in evaluation with which instructors and professional test writers struggle. This section is intended to identify and discuss those problems.

Validity

A working definition of validity is that a valid test measures what it is intended to measure effectively, efficiently, and consistently.

Evaluation as a sample. All evaluation is based on a sample of students' knowledge or performance. It is impossible to measure everything that students know or can do. It is impossible to create all of the various environments in which they will have to perform or anticipate all of the tasks they will be required to perform on the job. The problem then becomes one of identifying the knowledges or skills taught that are most likely to predict job success. The ultimate test of how well the sample of evaluation criteria was selected is whether successful students get jobs and perform satisfactorily.

There are a number of things that can be measured during the evaluation process. The objective is to measure knowledge, skills, and attitudes pertinent to good job performance. Such is not always the case and, thus, the validity is decreased. It is a question of determining what is really being measured. Many tests are actually measures of students' reading ability, memory, or fortitude. These are often related to successful job performance but are infrequently what was intended to be measured. It is unfortunate when such things as the students' knowledge of trivia, test-taking ability, or ability to out-guess the instructor are the major measure factors.

Difficulty. Another validity-related question is, how well must the student perform? In many instances individuals can agree on what knowledge and skills are necessary, or even more important, for successful job performance. However, those same individuals may have drastically different ideas about what constitutes reasonable competence. It is just as important that standards not be unrealistically

high or unrealistically low. Difficulty can be viewed in terms of the whole test or individual items on the test. Whether the test be valid or not, the difficulty of the instrument, as well as the items, must be appropriate to the purpose of the evaluation. Much time can be wasted in both instruction and evaluation by trying to achieve unrealistically high standards of performance. Since evaluation must be based on a sample of student performance, it is important that the sample be broad and representative.

Discrimination. The ultimate goal of evaluation is to discriminate between those students who have the knowledge and skills and those who do not. In order for a test to discriminate effectively, each item must discriminate. Thus it is very important that each item included in an evaluation instrument be constructed so that more knowledgeable or more skilled students earn higher scores. To include any items on which students with unsatisfactory levels of knowledge or skill regularly exceed the performance of students with satisfactory knowledge or skill reduces the validity of the evaluation.

Reliability

Reliability is generally associated with consistency. That is, can one count on the same thing happening again under the same circumstances? The reliability of a measurement instrument can be predicted by various statistical procedures, but reliability is really established over a period of time. It is important that a measurement instrument be reliable if it is to be useful. No test of students' knowledge or skill is "100 percent" reliable. The objective is to develop the best measuring instrument possible. Factors which generally affect the reliability of an instrument are (a) How adequately are the skills or knowledges sampled? (b) How clearly is what constitutes success defined? and (c) How appropriate is the difficulty of the test? All things being equal, tests that are relatively long, objective, and reasonably difficult will be more reliable than those that are relatively short, subjective, and too easy or too hard.

Evaluating Affective Skills

Affective skills are particularly difficult to measure and evaluate. Pencil and paper instruments are of limited value in most cases. If they are to be of any use, they must be carefully constructed since many experts believe that simply attempting to measure individuals' attitudes changes those attitudes. Thus instruments designed to measure attitudes directly are difficult for even professional test

developers to construct. It is for this reason that affective skills are usually measured indirectly.

Examples of the types of instruments used to measure affective skills indirectly are anecdotal records, rating scales, and checklists. These instruments are based on what the instructor observes. These can be observations of behavior or products. It should be apparent that validity and reliability are the major problems encountered in developing and using these types of instruments. Since they are indirect rather than direct measures, validity and reliability are very critical.

The validity and reliability of indirect measurement instruments can be improved by: (a) determining ahead of time what is important to observe, (b) recording observations immediately, (c) making a conscious effort to be objective, (d) recording factual descriptions, (e) using the instruments frequently, (f) clearly specifying directions and procedures to be followed, and (g) never evaluating students on the basis of one or two observations.

Making Evaluation Relevant

A number of questions that students and instructors raise regarding testing and evaluation were identified in a previous section of this monograph. The purpose of this section is to identify some procedures that instructors can use to solve problems posed by these questions.

Questions often arise about the relevancy of tests, test items, and the evaluation process. These questions are concerned with (a) what is evaluated, (b) the method of evaluation, and (c) the conditions under which evaluation takes place.

Selecting the Evaluation Sample

As has already been discussed, all evaluation is based on a sample of the students' knowledge or skill since it is impossible to measure everything students know, can do, or will be required to do. Thus it is very important that what will be evaluated is carefully selected.

Course objectives. The objectives of the course as reflected by the performance objectives should serve as the pool from which performances to be evaluated are selected. In some instances it may be possible, even necessary, to evaluate student performance on every performance objective. That still represents a sample of student performance. Evaluation is most often, but does not have to be, based on terminal performance objectives.

Although it is not necessary to evaluate student performance in terms of every course objective, it must be possible to do so. For example, affective skills are difficult to evaluate. However, if affective skill development is a course objective, instructors must be able to evaluate student affective performance. On the other hand, evaluations of students must **not** be based on knowledge or skills not included in course objectives and the associated performance objectives.

It can be difficult to determine exactly what a course objective encompasses. A competency-based approach to instruction encourages the development of clearly stated objectives. Well-defined objectives are one key to relevant evaluation.

Job requirements. Determining what students must know and be able to do if they are to obtain employment and be successful is the basis for developing instructional (enabling) objectives. Job requirements also provide insight into what should be evaluated. Since it is impossible to evaluate students in terms of everything required on the job, evaluated performances should be critical. There are three job-related criteria that are useful in selecting which performances to evaluate: (a) Is the knowledge or skill used frequently? (b) Is it difficult to master? and (c) Is it critical for entry-level employment? The major purpose is to separate the things students must be able to do from all others. Evaluation should focus on the necessities first. If time is available, students may be evaluated for basic knowledge and skills. Not included in the evaluation process are any advanced knowledge or skills. The efficiency of evaluation can be increased by concentrating on job-related tasks requiring students to have broad knowledge and be able to perform several skills.

Instruction. Instruction should be based on job requirements. However, everything required on the job is not taught in any single course. Moreover, no two classes progress at the same rate. In some cases, course objectives will be exceeded and in others all may not be met. Thus, it is inappropriate to evaluate all students and classes on exactly the same knowledge and skills. Evaluation should reflect instruction. Everything taught may not be evaluated, but anything not taught should not be evaluated. It is important that students be evaluated in terms of what they have learned in the course, not what they know beforehand. Whatever was emphasized during instruction should be emphasized during evaluation. Knowledge or skills taught for familiarization or introductory purposes should not be the focus of evaluation.

Establishing Realistic Criteria

Evaluation may be based upon a sample of students' knowledges and skills—appropriate in light of course objectives, job requirements, and instruction—and still not be relevant. Realistic evaluation criteria must be established and used. There are several factors that should be considered.

Course objectives. Just as they influence what is evaluated, course objectives affect what criteria are used. The performance objectives developed on the basis of the course objectives should specify very clearly what constitutes satisfactory performance. The criteria statement of a performance objective is almost always the most controversial part. Most instructors generally agree upon the performance and the conditions under which it must be performed; however, they often disagree upon what constitutes competence. Therefore, it is best if instructors maintain an open mind about the criteria they establish and validate them over a period of time. The main point is that knowledge and skill should never be evaluated by a higher standard than specified in the performance objectives.

Job requirements. Job requirements should closely relate to course objectives and should serve as the ultimate basis for establishing realistic evaluation criteria. The criteria used to evaluate students should conform with job requirements. They should be based on what is actually required, not what should be or what people say is required. Moreover, the criteria must be based on the requirements of the job for which the program is designed.

Apprentice masons and master masons must be able to perform many of the same skills. However, an apprentice is not required to perform with the same expertise as a master mason. If the program is designed to prepare apprentice masons, that is the level of expertise which should be established as the evaluation criteria. To do otherwise is similar to comparing a Ford Fiesta to a Rolls Royce. It can be done, but it doesn't tell you much if what you're really interested in is whether or not the Fiesta is a good small car.

Students' ability. The abilities of the students enrolled in the program should be considered when establishing evaluation criteria. There are individuals who argue that once standards for successful performance are established the students' ability levels are no longer a factor. That might be reasonable under two conditions: first, if we were absolutely positive of what actually constitutes competence; sec-

ond, if there were no related jobs that required a lesser degree of mastery. Neither case usually exists. Whether or not evaluation criteria are realistic in terms of students' ability provides a useful check of what were established as course objectives based on minimum job requirements. If standards seem too high in light of the students' abilities but realistic in terms of job requirements, some revision is in order. Additionally, since increasing numbers of special-needs students are entering occupational programs, in some instances it may be necessary to identify related jobs more in harmony with their abilities.

Time. Time or, more accurately, time limitations affect evaluation criteria in two ways. First is the question of whether or not time should be a part of the criteria. There are some tasks which must be performed within a certain amount of time. When this is the case, it is reasonable to include some time limitations in the evaluation criteria.

A second question to be considered is the amount of time devoted to evaluation. Evaluation standards must be reviewed in terms of whether or not it is realistic to expect students to achieve the standard in the time available and whether or not the skill or knowledge is critical enough to warrant the time required for evaluation.

Use Appropriate Techniques

If evaluation is to be relevant, techniques appropriate to what is being measured must be used. One of the first rules instructors teach their students is to always use the right tool for the job. That rule is just as valid in evaluating performance.

Skill evaluation. The evaluation of skills is best accomplished by using performance measures. There are some pencil and paper tests that approximate a performance test, but they are few. The ones that exist generally focus on process rather than product. When evaluating students' skills, the first thing to determine is whether or not it is important to evaluate the product, process, or both. If the product is of primary importance, the project method can be useful. However, if the process is either equally or primarily important, a traditional performance examination is most realistic. The amount of time required for evaluation is a major factor when both the product and process are important. In such cases, instructors should determine whether or not it is necessary that the entire project be completed and the entire process be observed.

Knowledge evaluation. Students' knowledge is most often

evaluated using pencil and paper instruments. However, oral exams and performance tests can also be used. Instructors should not overlook the possibility that standardized tests may be available and appropriate. There is no reason why standard measurements cannot be used in a competency-based setting if they are appropriate in terms of the course objectives, expected student outcomes, job requirements, and instruction. When constructing pencil and paper or analyzing standardized tests, it is important to insure that knowledge is evaluated at realistic levels. The limit on the level is defined by course objectives and job requirements. For example, it is unrealistic to expect auto mechanics to know the valve lash setting for all automobiles; however, they should be able to recognize reasonable clearances.

Environmental Considerations

Relevant evaluation takes place under realistic conditions and in an appropriate environment. Insofar as possible, evaluation, like instruction, should reflect the conditions that actually exist on the job. This is particularly true when evaluating skills. Few auto mechanics are ever required to trouble-shoot an engine that won't start when it is mounted on an engine stand. Therefore, trouble shooting skills should be evaluated using a complete automobile if at all possible. When using pencil and paper measures or other knowledge measures, instructors must insure that an appropriate environment exists. Distraction should be eliminated. The objective is to measure students' knowledge not their ability to work under adversity.

Making Evaluation Meaningful

If evaluation is to be meaningful, results must be used. Evaluation results can be used in a number of ways that make them meaningful to both students and instructors.

Students

The first principle of making evaluation meaningful to students is to provide feedback promptly. This is less of a problem in a competence-based program. Students are informed of expectations prior to instruction. Thus, when they are tested, they are often able to estimate how well they have performed. However, instructors should still make an effort to provide formal feedback as soon as possible after the evaluation process is complete.

A reasonable amount of time should be set aside to discuss the evaluation with students either as a group or individually. Discussing

why certain answers are right or wrong and analyzing what was good or bad about a procedure or product evaluation becomes a learning experience and reinforces what was taught. Allowing reasonable student input and criticism during this discussion decreases the anxiety associated with evaluation and creates a more positive learning environment.

The chance that students will view evaluation as a learning experience is also increased if the penalty for failure is minimized. One way to do this is to allow students to retake examinations if they perform unsatisfactorily. This practice is consistent with a competency-based program since the student's performance is evaluated against an established standard rather than how well other students performed. However, additional learning activities should usually be required between evaluations.

Instructors

Instructors can also find the results of evaluation to be meaningful and useful. The first thing that should be done is to evaluate the evaluation, either formally or informally. Student input often provides valuable insight into why certain things did or didn't work. The purpose is to reflect on the total evaluation process to determine in general how well it worked and which areas need improvement as well as how much faith to place in the outcomes.

Instructors should analyze students in light of their (instructor) performance evaluation. Such an analysis would (a) focus on determining why certain students performed as they did, (b) diagnose student weaknesses and strengths, and (c) determine how well the instruction met students' needs. If a majority of students performed poorly on an evaluation, it should signal that the instructor may have a problem. If the measurement instrument appears to be realistic, valid, and reliable, several alternative reasons should be evaluated. Perhaps the evaluation was conducted under adverse conditions or with faulty tools and equipment. Perhaps the instruction related to the particular objective was not as clear as originally thought. Analysis of this type can often point to a need to improve either what is taught, how it is taught, or how evaluations are conducted.

Analyzing an individual student's performance is useful in diagnosing strengths and weaknesses. This information can help instructors develop instructional plans that will be effective in meeting individual student's needs. Moreover, knowing a student's strengths and weaknesses can help instructors place their students more effec-

tively. Instructors can place students in jobs which offer the greatest chance for the student to succeed. This makes the instructor and the program look good while also making students and employers happy.

Making Evaluation Easier

A major concern of instructors is that evaluation is a time-consuming and laborious task. One of the reasons it seems so is that test planning and development is often postponed to the last minute. There are several ways in which this can be avoided.

Item or Problem Construction

One of the best ways of avoiding the pain often associated with constructing test items or problems is to do it ahead of time. It is generally best to develop several items or problems each day. This procedure spreads the task out over the entire school year and generally results in better tests. This is because the items or problems are developed while what was taught is still fresh in the instructor's mind and all instructional materials are readily available. A second procedure is to have students prepare items or problems as part of a review of units or blocks of instruction. With a little guidance, students can develop excellent test items. In addition to providing a good way for students to review, instructors can check students' understanding by checking and evaluating the items they develop. Care should be taken, however, to avoid using student-developed items or problems wholesale without evaluating their quality and appropriateness. Third, in many instances, it is possible to reword performance objectives and use them as test items or problems.

Two major guidelines should be followed in preparing test items or problems. First, each should be constructed with a specific purpose in mind; that is, it should be designed to measure a specific bit of knowledge or skill. Items and problems should be evaluated to determine if they are measuring what was intended or if they are really measuring reading ability, test-taking ability, memory, or something else not related to job requirements. Second, items or problems should deal with important knowledges or skills. If tests continually concentrate on measuring the trivial, students will concentrate on learning the trivial.

Instrument Construction

A key to making the construction of evaluation instruments easier

is developing a file of test items and problems. These can be written on note cards and stored according to course objectives or performance objectives. They can also be filed on the basis of subject matter or question/problem type. The card should include the question, the correct answer, procedure or criteria for success, and a notation indicating where explanatory information can be found. One of the easiest ways to develop a large file quickly is to trade test items or problems with other instructors in the field. Items or problems obtained from other instructors should, of course, be evaluated for appropriateness.

Once a reasonable file is developed, instrument construction becomes a matter of selecting appropriate items or problems from the file. The directions can then be written, the items or problems ordered, and the instrument typed from the cards. If equipment exists, it is possible to have each item or problem and appropriate directions punched on computer cards. These cards can be stored with the file cards. It is then possible to order the directions and item or problem cards, run them through a card reader, and have the instrument printed by a computer.

Regardless of the exact procedure used, the key is to develop a file of items and problems. This eliminates having to start from scratch every time. It also discourages the other alternative, using the same instrument over and over again. Students can always find a way to obtain a copy of the instrument if the instructor uses it repeatedly.

Correcting Instruments

When pencil and paper instruments are used, correcting them can become time-consuming. With a little planning before the instrument is constructed, considerable time can be saved. One alternative to correcting tests by hand is to have the students mark their answers on machine readable sheets and have them corrected by machine. If such a service is not available, time can be saved by developing overlay sheets. An overlay sheet is merely a sheet of paper that is laid on top of the instrument or answer sheet. Holes are punched where the correct answer is located. If a mark is visible in the hole, the student has answered correctly. If no mark is visible, an incorrect answer has been marked or the question has been skipped. In any case, a check is placed in the hole. The number of checks can be easily counted to determine the student's score. Additionally, when the instrument is returned to the student, the checks immediately indicate the correct answer. When an overlay sheet is going to be used to correct in-

struments, students must be instructed not to circle the correct answer but to mark an "x" through it or blacken it. Further, the student answer sheet should be scanned to be sure only correct answers are marked.

Another alternative exists and is applicable to performance tests of skills. Quite simply, let the student do it. Students can grade their own tests, or they can pass them around. In addition to the time saved, an advantage is that feedback is immediate. The biggest disadvantage is the possibility that some students may agree to change each other's answers. There are ways to guard against this possibility, such as having tests answered in pencil but graded with a ballpoint pen.

In the case of performance exams, students can work in teams, alternating between observer and examinee when it is important to evaluate the process. If this method is to be used, the instructor must develop a clear, concise, and explicit yes-no checklist and show students how to use it. This procedure is probably not effective if anything but a checklist is used. If the procedure is the focus of the evaluation, students can either evaluate their own products or the products can be shuffled. Again, unless the majority of the evaluation criteria can be specified on a yes-no checklist, the procedure is of questionable value.

Improving Evaluation

The main principle that must be followed if evaluation is to be improved is to plan the evaluation process. This plan may not be written, but at least it should be mentally developed. The purpose of the evaluation, the material to be covered, the relative importance of the areas to be covered, and the criteria that will indicate satisfactory performance should be indicated on this plan. Once the plan is developed, the measurement instrument can be developed.

Pitfalls to Avoid

There are common pitfalls in item or problem construction that instructors should avoid. It is often surprising that students are more aware of these than are instructors. By knowing them, students are able to raise their performance level without knowing answers or having skills. Some examples include:

Identifiable answer patterns. Some answers follow a specific format; i.,e.,A,D,B,C,A,D,B,C.

Specific determiners. True-false items containing words such as

always and **never** are always false; conversely, those containing words like **may** and **sometimes** are almost always true.

Longest alternatives. When answering true-false items, long items are more often true than false. When answering multiple-choice items, the longest alternative has the greatest chance of being correct.

Verbal association. Sometimes it is possible to identify the correct alternative in multiple choice items because a word in the alternative is related to a word in the stem.

Overlapping distracters. In multiple choice items if two alternatives are worded so that each makes the other correct, they can be disregarded.

Grammatical clues. It is sometimes possible to identify the correct alternative in multiple choice items because some alternatives cannot be added to the stem and be grammatically correct. This often occurs when the stem ends in words such as **are, is, a** and **an.** Other examples may occur under other conditions.

Skill sequence. In a performance exam if tools are arranged in the order they are used, students may be reminded of or able to guess the next step.

Directions. Just as it is important to develop clear, concise, unambiguous test items or problems, it is important that students understand what is to be done. The importance of writing good directions can hardly be over-emphasized. Unless students know exactly what is to be done, the time and effort spent in developing good items or problems is wasted. Poor directions present the student with a very frustrating problem, quite similar to trying to assemble a bicycle on Christmas Eve! When inadequate directions are provided, the students' ability to guess what is desired is more accurately measured than their knowledge or skill. One good way to avoid this problem is to write down the directions and ask students to read them and interpret what they mean. If having students in the class read them will compromise the evaluation, use similar students from another class.

Variety. It is important to avoid using only one or two types of items, problems or approaches to evaluation. Using a variety gives a more accurate measure of what the student knows or can do. Just as different students learn in different ways, students have preferences in types of items or problems. For example, more knowledgeable students often have more difficulty with true-false items than less knowledgeable students.

Time. Unless time is to be included as one of the criteria for success, every effort should be made to eliminate time as a factor. If the objective is to measure what the students know or can do, then the instrument should be conducted so that almost all students can complete it in the time available. It is always difficult to know how long students will need. An estimate can be reached by having the instructor complete the task and then multiply the time it took by two. The need to allow adequate time so that the majority of students can complete the task increases the importance of having a definite plan for selecting what is important to measure and how to measure it.

Formal Instrument Evaluation

After an evaluation has been conducted, it is often useful to somewhat formally evaluate the instrument and the items or problems. This information can be useful in analyzing students' performance as well as improving the items or problems in the evaluation file. The results also provide information about how accurately knowledge or skills were measured and how much faith should be placed on the outcome.

Validity. The validity of the instrument and of individual items or problems should be checked before the instrument is used. Often, however, things which appear valid to the instructor do not appear so to students. Therefore when problems arise, instructors should seek student input as well as re-evaluate their own assessment. In terms of evaluating the measurement instrument in general, the following questions should be asked:

1. Is what was taught adequately covered?
2. Are the knowledge and skills emphasized in instruction emphasized to an equal degree in the instrument?
3. Does the instrument assess important knowledge and skills based on job requirements?
4. Are the directions clear, concise, and unambiguous?

Difficulty. Whether or not a performance exam is of appropriate difficulty cannot be computed but is a judgment made by the instructor. The difficulty of a performance test is established based upon the difficulty of the task involved and the criterion statement in the performance objective. For performance tests, it is important that the difficulty of the test be consistent with the difficulty of the task. If the task is easy, the test should be easy. Any attempt to make

the test more difficult will usually be based on tricks to fool the student. On the other hand, if the task is difficult, the test should be appropriately difficult.

The difficulty of the pencil and paper test is determined by the difficulty of its components and can be computed. The difficulty of an item is determined by dividing the number of students answering correctly by the total number of students taking the exam. What constitutes an appropriate level of difficulty depends on the purpose of the evaluation. If the test is designed to rank students in order of their knowledge, items of approximately 50% difficulty should be used. If the purpose is to diagnose student strengths and weaknesses, items of varying difficulty are appropriate. The difficulty of items or problems can be established over a period of time and should be noted on the card containing the item or problem in the test file.

Discrimination. The discrimination power of an item is its ability to separate those who know more from those who know less. However, in a criterion-referenced evaluation the critical issue is whether or not the item measures the effects of the instruction. In a traditional pencil and paper measurement setting, the objective is to construct items or problems in such a manner that students scoring well on the exam are more likely to get the item correct than those scoring poorly. This can easily be checked. Divide the students into groups according to their test scores. Compare the responses of the top 25 percent to the bottom 25 percent for each item. Any item which students in the bottom group answer correctly more often than those in the top group should generally be rewritten or eliminated. An index of discrimination can be computed by subtracting the number of students in the low group getting the correct answer from those in the high group getting it correct and dividing by the total number of students in both groups. The resulting value can range from -1.00 to +1.00. Those items which have the best discrimination power can be identified by indexes close to +1.00. The discrimination index should also be entered on the card containing the item or problem in the test item file.

When performance evaluations are used, test items or problems should establish what a student has learned during instruction. This is discrimination power. Thus it is necessary to compare a pre-test to a post-test. The discrimination index can be computed by subtracting the number of students correctly answering the item/problem or completing the task before instruction from those who were suc-

cessful after instruction and dividing by the total number of students. This value can also range from -1.00 to + 1.00, and appropriate items will have indices greater than zero. Items, problems,and tasks having indices near zero or below should be revised or eliminated.

Evaluation and Grades

The fact that grades are important cannot be denied. It is also fair to say that both students and instructors are concerned about grades. In some instances grades appear to have become more important than learning.

It is important that grades be based upon some form of measurement. One purpose of evaluation is to develop information for assigning grades. How fairly grades are assigned depends upon relevant, accurate measurement. There are a number of reasons given for assigning grades, and some are more controversial than others. Whether the examples listed below seem reasonable will depend on how you view the purpose and process of evaluation. Thus, they are listed without comment.

1. Admit or exclude individuals from educational and occupational opportunities.
2. Help students identify their strengths and weaknesses.
3. Inform students of their progress.
4. Motivate students to learn.
5. Describe a student's level of competence.
6. Evaluate instruction.

Traditional grading is normative based. That means that each student's performance is compared to the performance of all other class members. Since no two classes are exactly equal in ability, knowledge, interest, or other characteristics, there is no solid base from which to gauge what grades assigned in this manner really mean.

It is often a fact of life that students must be assigned traditional **A** to **F** grades. This can be a problem for instructors whose programs are competency-based. Several strategies can be used to convert competency-based evaluations to traditional grades.

1. Grades can be assigned on the basis of the number of competencies a student is able to learn during the marking period.
2. Competencies can be evaluated and weighted according to their importance for job entry. Thus, rather than merely counting the number of competencies the student can demonstrate, the importance of these competencies is considered.
3. Criteria can be established for assigning a grade to varying

levels of success in demonstrating each competence. For example, minimum competence as defined by job entry requirements might be assigned to a grade of "C." Then the quality of the work would be evaluated to determine (a) how much it exceeded the minimum competency level for assigning grades of **A** or **B**, or (b) how much it fell below the minimum competency level for assigning grades of **D** or **F**.

4. Students can write a contract with the instructor for various grades. This contract should specify what the student will do, how well it must be done, and how it will be evaluated for an agreed grade.

If traditional grades must be assigned to students enrolled in a competency-based program, the system to be used must be defined. Moreover, students should be informed about how this will be accomplished.

Alternative to Grades

Some alternatives to traditional grades exist. Instructors may find the first two to be useful in conjunction with traditional grades. The first, and probably most useful to occupational instructors, is competency checklists. Competency checklists include all of the competencies taught in a particular course. They can be constructed so that the instructor merely checks off the items when the student successfully demonstrates competence. Some checklists may require the instructor to rate the student's level of competence. Competency checklists are particularly useful to employers in evaluating a student's qualification.

Written evaluation reports are another alternative. In actuality, most instructors are using this alternative but are calling it a personal recommendation. It is just a factual description of the student's strengths and weaknesses.

Another alternative is to replace traditional grades with either pass/fail or satisfactory/unsatisfactory grades. A compromise would be to allow students to choose between receiving a traditional grade and a pass/fail or satisfactory/unsatisfactory grade. If either the pass/fail or satisfactory/unsatisfactory system is used, competency checklists or written evaluations are almost always required if grades are to provide information to students, parents, or prospective employers.

Annotated Outreach

A large number of books, modules, paperbacks, and articles have

been written on evaluation, testing, test construction, and related issues. Instructors should find any of the following easy to obtain. Additionally, they are written in a manner that makes them easily understood and useful.

The Center for Vocational Education. **Instructional Evaluation Professional Teacher Education Module Series - Category D.** Athens, Georgia: AVVIM, 1978.

The six performance-based modules are entitled (a) **Establish Student Performance Criteria;** (b) **Assess Student Performance: Knowledge;** (c) **Assess Student Performance: Attitudes;** (d) **Assess Student Performance: Skills;** (e) **Determine Student Grades;** and (f) **Evaluate Your Instructional Effectiveness.**

Gronlund, N. E. **Measurement and Evaluation in Teaching** (3rd Edition). New York: MacMillan Publishing Co., Inc., 1976.

This text gives extensive coverage to almost all aspects of evaluation. The discussion is complete and understandable. Many practical examples are used. Major points, rules, or hints are included and discussed in almost every section, making it a useful resource book. Major issues are discussed as they apply to criterion-referenced as well as norm-referenced evaluation.

Smith, F. M. and Adams, S. **Educational Measurement for the Classroom Teacher.** (2nd Edition) New York: Harper and Row, Publishers, 1972.

A basic book on the construction and use of teacher-made and standardized tests. Overall, it is understandable with the exception of the chapter on statistics. Each chapter contains an annotated bibliography of additional readings. Major points are illustrated using good examples.

Sign Off

Instructors and students are concerned about evaluation. Many of their concerns are similar. Both want evaluation to be fair, reliable, accurate, and relevant. Additionally, both tend to see evaluation as being a time-consuming process.

A number of suggestions have been offered regarding how to:

1. Make evaluation relevant by basing it on course objectives, job requirements and instruction; establishing realistic criteria; using appropriate techniques; and establishing a realistic environment.
2. Make evaluation meaningful by identifying how results can be useful to students and instructors.
3. Make evaluation easier by identifying efficient procedures for

item or problem construction, instrument construction, and correcting instruments.

4. Improve evaluation by following a defined plan, avoiding common measurement problems, and evaluating the evaluation.

5. Use evaluation in assigning grades including (a) converting criterion-referenced grades to traditional grades and (b) identifying alternatives to grades.

The evaluation process can be useful in (a) helping students to learn efficiently, (b) improving instruction, and (c) justifying the program improvement. The following checklist should serve as an efficient device for diagnosis and prognosis of your evaluation approach. Try to overcome the "specific determiners." By and large, the questions in the "before" and "during" sections should by "yes"; the questions in the "after" section are prescriptive. Please note that each of the questions on the checklist has been addressed in the monograph.

Checklist for Evaluating Evaluation

Before the evaluation	Yes	No
1. Was evaluation developed according to a plan?		
2. Is the purpose of the evaluation defined?		
3. Are all important knowledges and skills measured?		
4. Is the evaluation consistent with what was taught?		
5. Has subjectivity been minimized?		
6. Are criteria clearly specified?		
7. Are criteria realistic?		
8. Are appropriate measurement techniques used?		
9. Can most students finish in the time available?		
10. Can the results be used for purposes other than assigning grades?		
11. Can scoring be done by students or computer?		
12. Have plans been made to review results with students?		
13. Are directions clear and unambiguous?		
14. Is a variety of item or problem types used?		
15. Does each item have a definite purpose?		
16. Have items been checked for common errors?		
During the evaluation		
1. Was the evaluation made under appropriate conditions?		

	Yes	No
2. Did equipment function properly?		
3. Were the necessary supplies, tools, and equipment available?		
4. Did the exam proceed without interruptions?		
5. Did students proceed without misunderstanding?		
After the evaluation		
1. Did students perform as expected?		
2. Are students weak in some areas?		
3. Is additional instruction needed?		
4. Should some instruction be repeated?		
5. Do students need additional practice?		
6. Are changes in instructional techniques needed?		
7. Was each item or problem evaluated?		
8. Was student input sought?		
9. Were results reviewed with students?		

ERIC Clearinghouse for Junior Colleges
96 Powell Library Building
University of California
Los Angeles, California 90024

JAN 16 1980